Project: Capstone Project 2: Milestone Report 2

Machine Learning:

In the final phase after data pre-processing, to prepare our data for modeling, it was imporatnt to check for missing and categorical values and see if it requires any further treatment.

There were still multiple missing values present in the dataset which had to be topped up by frequently occurring values a.k.a mode values in the column or replaced with 'Unknown' where mode values presented a bias due to large amount of missing information.

A total of **155,047** values in various columns were replaced to completely avoid missing fields in the final table, df_cleaned.

In the last step before modeling, multiple categorical values had to be converted in to numerical values, which was done by using category encoding method discussed below:

1. One-hot-encoder.
2. Binary encoder.

The main difference between the two approaches was in terms of the data size it generated after conversion.

For example: One-Hot generated 3,912 columns after optimizing with label encoder while Binary Encoder generated 136 columns, which was significantly less when compared to one-hot ecoder.

In order to evaluate the difference between the two approaches, I modeled two data sets separately, Model A (one-hot) and Model B (Binary) before applying machine learning model.

Model Selection:

Given the data structure, I decided to evaluate the two models using Random Forest Classifier. Random forests can handle unbalanced data, outliers and non-linearity well. The only drawback however is that it can overfit the data which can be managed by hyperparameter tuning.

To identify best hyper parameters for the model, I started with Randomized Search CV and finished selection using Grid Search CV. During this process, we made 90 fits and it took 1 hour and 9 minutes to get the following hyper parameters using Model B data.

Selected Hyper Parameters	Description
{'n_estimators': 100,	#Number of Trees
'min_samples_split': 10,	#Minimum samples required to spilt internal node
'min_samples_leaf': 1,	#Minimum samples at leaf node
'max_features': 100,	#Maximum features to consider for split decision
'max_depth': 50,	#Maximum depth of each tree
'criterion': 'entropy'}	#Quality of split

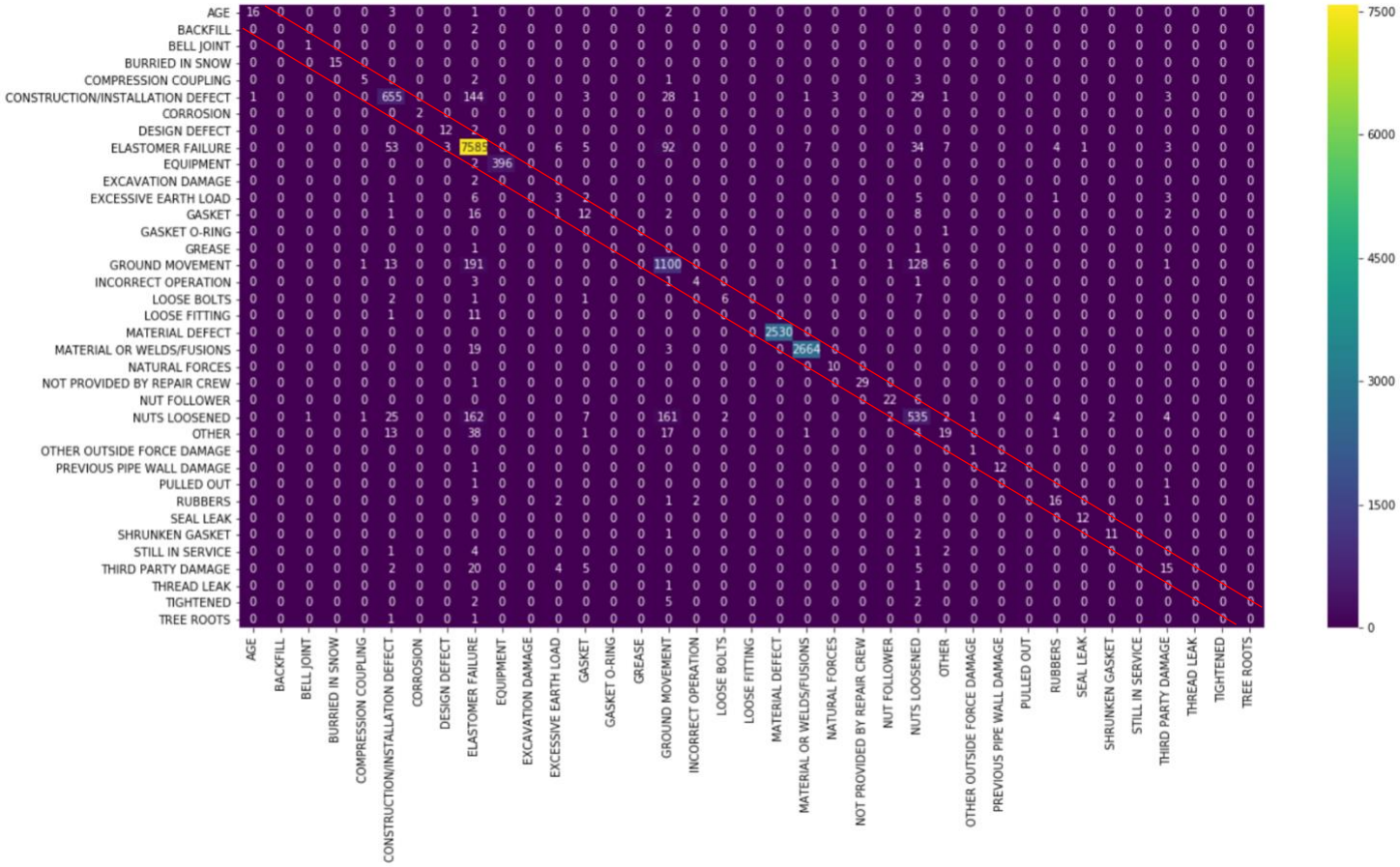
Model Evaluation:

Model B (Binary Encoded): As per classification report, we were able to train and classify 38 reasons that caused a mechanical fitting leak in the gas pipeline. Below is the classification report and a visual of confusion matrix, which will simply tell us how many correct and incorrect predictions for each leak label was made .

Classification Report:

	precision	recall	f1-score	support
AGE	0.94	0.73	0.82	22
BACKFILL	0.00	0.00	0.00	2
BELL JOINT	0.50	1.00	0.67	1
BURRIED IN SNOW	1.00	1.00	1.00	15
COMPRESSION COUPLING	0.71	0.45	0.56	11
CONSTRUCTION/INSTALLATION DEFECT	0.85	0.75	0.80	869
CORROSION	1.00	1.00	1.00	2
DESIGN DEFECT	0.80	0.86	0.83	14
ELASTOMER FAILURE	0.92	0.97	0.95	7800
EQUIPMENT	1.00	0.99	1.00	398
EXCAVATION DAMAGE	0.00	0.00	0.00	2
EXCESSIVE EARTH LOAD	0.19	0.14	0.16	21
GASKET	0.33	0.29	0.31	42
GASKET O-RING	0.00	0.00	0.00	1
GREASE	0.00	0.00	0.00	2
GROUND MOVEMENT	0.78	0.76	0.77	1442
INCORRECT OPERATION	0.57	0.44	0.50	9
LOOSE BOLTS	0.75	0.35	0.48	17
LOOSE FITTING	0.00	0.00	0.00	12
MATERIAL DEFECT	1.00	1.00	1.00	2530
MATERIAL OR WELDS/FUSIONS	1.00	0.99	0.99	2686
NATURAL FORCES	0.71	1.00	0.83	10
NOT PROVIDED BY REPAIR CREW	1.00	0.97	0.98	30
NUT FOLLOWER	0.88	0.79	0.83	28
NUTS LOOSENED	0.69	0.59	0.63	909
OTHER	0.50	0.20	0.29	94
OTHER OUTSIDE FORCE DAMAGE	0.50	1.00	0.67	1
PREVIOUS PIPE WALL DAMAGE	1.00	0.92	0.96	13
PULLED OUT	0.00	0.00	0.00	3
RUBBERS	0.62	0.41	0.49	39
SEAL LEAK	0.92	1.00	0.96	12
SHRUNKEN GASKET	0.85	0.79	0.81	14
STILL IN SERVICE	0.00	0.00	0.00	8
THIRD PARTY DAMAGE	0.45	0.29	0.36	51
THREAD LEAK	0.00	0.00	0.00	2
TIGHTENED	0.00	0.00	0.00	9
TREE ROOTS	0.00	0.00	0.00	2
micro avg	0.92	0.92	0.92	17123
macro avg	0.55	0.53	0.53	17123
weighted avg	0.91	0.92	0.91	17123

Confusion Matrix:



Confusion matrix shows, counts outside of the two diagonal lines as mixed classification labels.

Feature Importance:

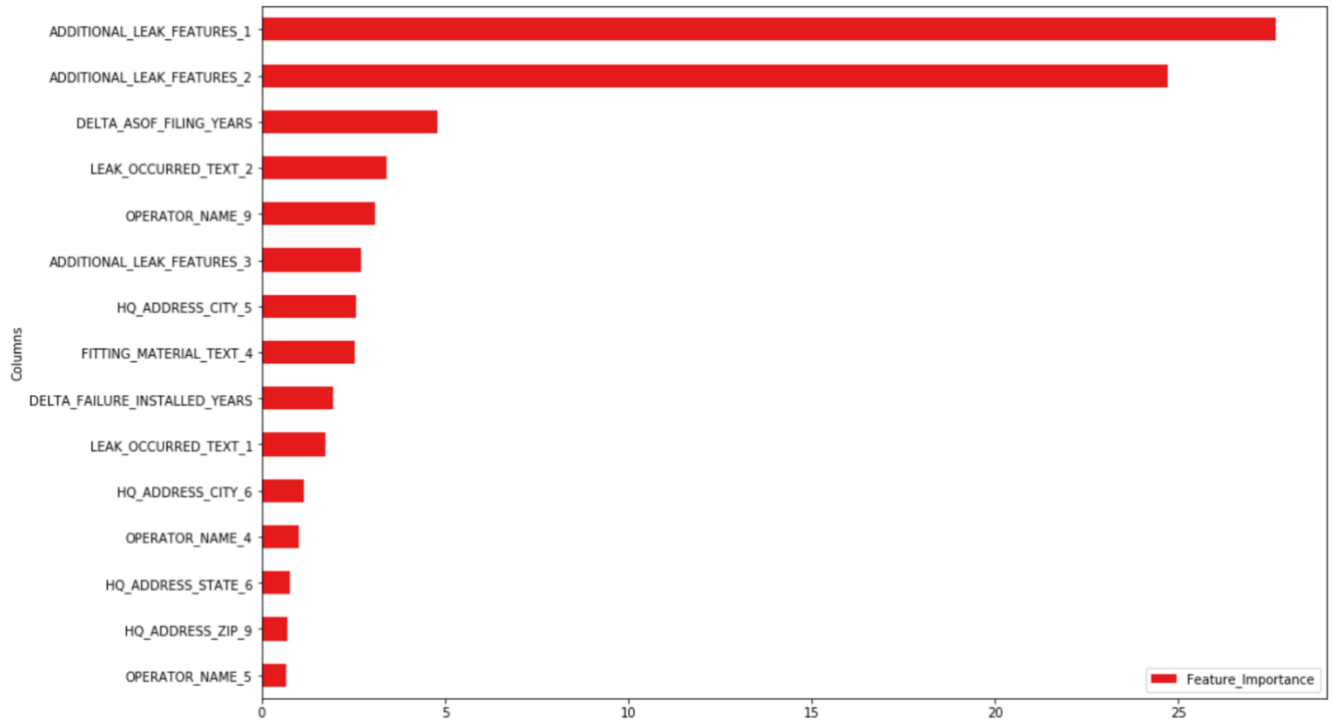
Classifier shows top 15 important features which contributed 79.9% in predicting leak cause in the pipeline with 91.6 accuracy.

```
Top 15 Features for Classifying Leaks in Pipeline:
 Feature_Importance    79.525348
dtype: float64
```

Columns	Feature_Importance
ADDITIONAL_LEAK_FEATURES_1	27.665679
ADDITIONAL_LEAK_FEATURES_2	24.703340
DELTA_ASOF_FILING_YEARS	4.783924
LEAK_OCCURRED_TEXT_2	3.396982
OPERATOR_NAME_9	3.096549
ADDITIONAL_LEAK_FEATURES_3	2.719892
HQ_ADDRESS_CITY_5	2.588522
FITTING_MATERIAL_TEXT_4	2.554088
DELTA_FAILURE_INSTALLED_YEARS	1.934990
LEAK_OCCURRED_TEXT_1	1.751176
HQ_ADDRESS_CITY_6	1.153430
OPERATOR_NAME_4	1.029005
HQ_ADDRESS_STATE_6	0.767364
HQ_ADDRESS_ZIP_9	0.716673
OPERATOR_NAME_5	0.663734

Distribution of top 15 features used in classifying pipe leaks:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8403617240>
```



Model A (one-hot):

Using same hyper-parameters, Model A resulted in comparable model accuracy 90.4% and leak cause classification prediction. However, it took longer run time and required way more more features compared to Model B.

Conclusion:

- Random Forest using Binary Encoded Data (Model B) classified top 15 causes for leak with 91.6% accuracy.
- Random Forest using one-hot encoded data (Model A) classified top 15 causes for leak with 90% accuracy.
- Model B run time was faster than Model A due to significant difference in column sizes (136 Vs 3,912)
- Model B used top 15 features to predict 79.9% of the leak classification while Model A classified only 48.8%.
- Random Forest using Binary Encoder (Model B) was preferred.
- Important Features which classified leak cause correctly were:
  - Excavation damages
  - Natural forces
  - Welding defects
  - Leak Through the Body or Seal
  - Time elapsed between 'Installation' and 'Failure Date'.

- Time elapsed between 'As of' and 'Filing Date'
- Pipeline Operator

Future Improvements:

Existing Model: Even though Model B accuracy score is 91.6%, there were areas of mixed classification which could be improved in future models.

> Example: Elastomer Failure had 7,642 examples where Leak was identified due to elastomer failure. However, it also had 144 examples of construction/installation defects, 191 examples of ground movement defects and 162 examples of Nuts Loosened defects under the same class.
>
> We would need to check with SME's to see if elastomer failure could be caused due to these examples.
>
> Moreover, exploratory analysis showed significant reasons for leak caused due to equipment failure. We would need equipment data to understand which equipment features led to fitting failure in the past, which is currently not known.

Apply Other Advanced Models: I would like to run deep learning model on the same dataset and compare leak classification result. Also, use predictive modeling to estimate when a mechanical fitting could fail based on the number of years of operations.

Build Dashboard: I would like to build a real-time dashboard on key performance indicators such as expected failure date, operator and manufacturer details, and likely reason for leak cause, to help clients avoid any operational risks, early-on.